

به نام خدا



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

توسعه سیستم پیشنهاددهنده کلیدواژه برای مستندات علمی فارسی بر پایه اطلاعات اصطلاحنامه‌های تخصصی

تاریخ شروع طرح پژوهشی	۱۴۰۰/۶/۲۲	تاریخ پایان طرح پژوهشی	۱۴۰۲/۷/۲۳
نام و نام خانوادگی مجری طرح پژوهشی	آزاده محبی، مرضیه زرین‌بال ماسوله		
نام و نام خانوادگی همکاران طرح پژوهشی	زهرا دهسرای، امیر بادامچی، امیر محمد همتی، محمد صدرا نجاتی		

درباره طرح پژوهشی

کلیدواژه‌های یک سند علمی مانند پایان‌نامه‌های موجود در پایگاه اطلاعات علمی ایران (گنج)، بیانگر موضوعات نهفته در سند است که در سیستم‌های بازیابی اطلاعات برای افزایش دقت روش‌های بازیابی، در نظر گرفته می‌شوند. علاوه بر بهبود بازیابی اطلاعات، کلیدواژه‌ها می‌توانند در خلاصه‌سازی خودکار، دسته‌بندی، خوشه‌بندی و نیز بازنمایی فراداده‌های یک سند برای مدیریت بهتر اسناد بکار گرفته شوند. در این طرح پژوهشی، سیستمی طراحی و توسعه یافته که می‌تواند بر مبنای عنوان و چکیده یک سند، کلیدواژه‌های مناسب برای آن را پیشنهاد دهد. روش‌های بدون ناظر برای استخراج خودکار کلیدواژه، معمولاً مجموعه از واژگان یا عبارات را براساس ویژگی‌های مختلف از متن به عنوان عبارات کاندید استخراج می‌کنند و با روش‌های امتیازدهی، از بین آنها بهترین عبارات به عنوان کلیدواژه انتخاب می‌شوند. در بسیاری از این روش‌ها، فراوانی به عنوان یکی از مهمترین ویژگی‌های عبارات کاندید در نظر گرفته می‌شود. از آنجاییکه در این پژوهش تنها به عنوان و چکیده اکتفا می‌شود، فراوانی لزوماً نمی‌تواند معیار مناسبی برای شناسایی کلیدواژه‌ها باشد و به دلیل کوتاه بودن سند، سایر ویژگی‌ها نیز کارایی مناسبی نخواهد داشت. هدف اصلی از انجام این پژوهش، توسعه سیستم پیشنهاددهنده‌ای است که بتواند با استفاده از منابع دانش بیرونی از جمله اصطلاحنامه‌های تخصصی، برای اسناد علمی مانند پایان‌نامه‌ها و رساله‌های پایگاه گنج ایرانداک، کلیدواژه‌های مناسبی را پیشنهاد دهد. روش پیشنهادی از دو مرحله تشکیل شده است. برای یک سند ورودی، در مرحله اول، از بین اسناد مشابه که قبلاً نمایه شده‌اند، کلیدواژه‌هایی از آن اسناد که با سند مورد نظر نیز مرتبط هستند، استخراج می‌شوند. سپس در مرحله دوم، با استفاده از اصطلاحنامه‌های تخصصی و متن سند مورد نظر، کلیدواژه‌های جدیدی نیز به کلیدواژه‌های مرحله قبل اضافه می‌شوند و در نهایت بهترین کلیدواژه‌های از نظر میزان ارتباط با سند پیشنهاد می‌شوند. برای ارزیابی عملکرد روش پیشنهادی از مجموعه داده حاصل از پژوهش محبی و همکاران (۱۴۰۰) استفاده شده است و نتایج آن با شاخص‌های

روش پژوهش

برای پیاده‌سازی سیستم پیشنهادی، دو گام اصلی در نظر گرفته شده است: (۱) طراحی و ارزیابی روش پیشنهادی استخراج کلیدواژه، و (۲) طراحی و پیاده‌سازی سیستم پیشنهاد کلیدواژه براساس روش پیشنهادی در گام نخست.

در گام نخست، روشی طراحی و پیاده‌سازی شده که با دریافت اطلاعات یک سند مانند عنوان و چکیده و موضوع، به عنوان خروجی، مجموعه‌ای از کلیدواژه‌ها را برای آن سند پیشنهاد می‌دهد. روش پیشنهادی علاوه بر تحلیل محتوی سند، با دسترسی به اطلاعات اسناد مشابه به سند ورودی (مانند عنوان، چکیده، کلیدواژه‌ها و موضوع)، و واژگان مرتبط با سند از اصطلاحنامه‌های تخصصی، کلیدواژه‌هایی را ارائه می‌دهد. برای بازیابی اسناد مشابه، از پایگاه گنج استفاده شده که سیستم بازیابی اطلاعاتی برای بازیابی اسناد مشابه از گنج طراحی و پیاده‌سازی شده است. سیستم طراحی شده، برای بازیابی اسناد مشابه از مدل‌های زبانی مانند word2vec و paraphrase-multilingual-mpnet-base-v2 استفاده می‌کند. برای دستیابی به واژگان از اصطلاحنامه‌ها از API اصطلاحنامه‌های ایرانداک استفاده شده است.

در گام دوم، سیستم پیشنهادی براساس روش طراحی شده در گام نخست طراحی و پیاده‌سازی شده و برای ارزیابی روش نیز از مجموعه داده‌های استاندارد متشکل از چکیده‌های ۲۸۳۸ پایان‌نامه و رساله، پیاده‌سازی شده و نتایج آن براساس شاخص‌های مختلفی نظیر دقت، بازخوانی و AP@K ارزیابی شده است استفاده شده است.

یافته‌های طرح پژوهشی

یافته‌های این پژوهش عبارتند از:

- از بین سه مدل زبانی مختلف ارزیابی شده در این پژوهش روی مجموعه‌ای از ۱۰۰ هزار داده مستخرج از پایگاه گنج، مدل paraphrase-multilingual-mpnet-base-v2 بهتر از بقیه می‌تواند شباهت بین متون را شناسایی کند و به عنوان یکی از مدل‌های مورد استفاده برای محاسبه شباهت بین متون و کلمات استفاده شده است.
- روش پیشنهادی از نظر دقت و فراخوانی و AP@K از چهار روش مورد بحث در این طرح، بهتر عمل می‌کند و نتایج معنادارتری را ارائه می‌کند.
- در صورتیکه در مجموعه داده‌های موجود برای بازیابی اسناد مشابه، در یک حوزه موضوعی خاص، تعداد محدودی سند موجود باشد، روش پیشنهادی در شناسایی و استخراج کلیدواژه دقت و بازخوانی پایین‌تری را خواهد داشت.
- از آنجاییکه در روش پیشنهادی از واژگان و عباراتی که قبلاً به عنوان کلیدواژه براساس اسناد دیگر در نظر گرفته شده‌اند یا از واژگان یا عبارات موجود در اصطلاحنامه‌های تخصصی استفاده می‌کند، نمونه‌های پیشنهاد شده برای کلیدواژه، خطاهای نگارشی و ساختاری حداقلی را نسبت به روش‌هایی که مستقیماً از متن کلیدواژه استخراج می‌کنند، دارند. در واقع در روش‌هایی که مستقیماً براساس واژگان مستخرج از متن سند کار می‌کنند، کلیدواژه‌ها براساس شاخص‌هایی نظیر فراوانی و هم‌رخدادی با سایر واژگان به عنوان کلیدواژه پیشنهاد می‌شوند. در این روش‌ها ممکن است ساختار و نحوه نگارش واژه پیشنهادی با استانداردهای معمول که برای ساختار کلیدواژه‌ها در نظر گرفته می‌شود، متفاوت باشد.

- هر چه روش بازیابی اسناد مشابه بهتر عمل کند و بهتر بتواند اسناد مشابه را استخراج کند، نتیجه نهایی بهتر خواهد بود.

نتیجه گیری و پیشنهادها

برای تحقیقات آتی می توان در هر یک از مراحل روش پیشنهادی بهبودهایی انجام داد تا در نهایت دقت و فراخوانی روش پیشنهادی افزایش پیدا کند. در مرحله اول، می توان از روش های مبتنی بر یادگیری عمیق استفاده کرد. در مرحله دوم هنگام غنی سازی واژگان بر حسب اینکه واژه یافت شده در اصطلاحنامه جزو واژگان وابسته، اعم یا اخص بوده، می توان وزن های متفاوتی برای آن در نظر گرفت، یا وزن ها را در محاسبه امتیاز نهایی آن لحاظ کرد. همچنین می توان از سایر روش های معمول استخراج کلیدواژه مبتنی بر گراف مانند روش SCAKE، یا در صورت موجود بودن امکانات پردازشی از روش های مبتنی بر یادگیری عمیق مانند FRAKE استفاده کرد. این روش ها می توانند در کنار روش پیشنهادی در این پژوهش برای غنی سازی فهرست کلیدواژه ها استفاده شوند.